# Measuring and Predicting Importance of Objects in Our Visual World

Computation and Neural Systems Technical Report CNS-TR-2007-002
http://resolver.caltech.edu/CaltechAUTHORS:CNS-TR-2007-002

Merrielle Spain and Pietro Perona
California Institute of Technology
Pasadena, CA 91125
{spain, perona}@caltech.edu

## Abstract

*Associating keywords with images automatically is an approachable and useful goal for visual recognition researchers. Keywords are distinctive and informative objects. We argue that keywords need to be sorted by 'importance', which we define as the probability of being mentioned first by an observer. We propose a method for measuring the 'importance' of words using the object labels that multiple human observers give an everyday scene photograph. We model object naming as drawing balls from an urn, and fit this model to estimate 'importance'; this combines order and frequency, enabling precise prediction under limited human labeling. We explore the relationship between the importance of an object in a particular image and the area, centrality, and saliency of the corresponding image patches. Furthermore, our data shows that many words are associated with even simple environments, and that few frequently appearing objects are shared across environments.*

Figure 1. A useful description for this image is a list of keywords such as 'pancakes', 'cantaloupe', 'breakfast', 'fork' , 'knife'. Keywords refer to important objects in the scene, i.e. objects that are distinctive and informative. (Photograph by Stephen Shore)

## 1. Introduction

'Image understanding', the grand goal of machine vision, is about computing meaningful and informative semantic scene descriptions from images. Consider the picture shown in Fig 1: You are about to have a pancake breakfast in a '70s restaurant in the Western US. The cantaloupe looks fresh and inviting, a glass of milk is filled to the brim. Quick, get started before your pancakes get cold! In order to sit down you need to pull back the red chair... Producing a description (or its equivalent in machine-understandable form) accounting for objects, materials, relationships, and actions is the long-term goal of visual recognition researchers.

Progress in visual recognition has been breathtaking during the past 10 years. We now have algorithms that can recognize individual objects accurately and quickly [10], algorithms that can detect categories of objects in clutter [4], classify scenes [11], learn new categories with little supervision [17, 5, 13, 6, 8]. Some of these algorithms are fast [14, 10, 6, 8]. A state-of-the-art algorithm can classify single object Caltech-101 images, producing one descriptive word per image [9].

What are the next steps toward image understanding? A full description of complex scenes, such as the one shown in Fig. 1, currently appears to be out of reach (although there is interesting work in that direction [2]). An intermediate goal is generating a list of keywords for a picture. This simpler description would be useful for indexing into large image databases (think of the keyword system in flickr.com) and it would be readily understandable by humans. How should such a list be produced? As we shall see later, images such as the one in Fig. 1 contain an enormous number

of objects (we will be using the word 'object' to cover visual phenomena that may be labeled by a word, e.g. 'frog', 'leather', and 'shadow'). Rattling off an alphabetized laundry list of nouns would not be particularly informative — not all words are equal. For instance 'pancakes' appears to be more informative than 'spoon' in Fig. 1. Keywords are salient and distinctive objects in the scene — 'important' objects. So our goal is to produce a *prioritized* list of the *important* objects in the scene. We formalize this concept as

> An object's *importance* is the probability that it will be mentioned first by an observer.

This study is about defining, measuring, and predicting the importance of objects in images. In section 2 we describe a a novel method for collecting keyword statistics from human observers. Section 3 quantifies the object richness of images, environments, and the world. To estimate importance, section 4 introduces a model for object naming based on importance. We show that this model accounts for both object naming frequency and order. We provide a method for measuring importance using this model. Section 5 attempts to predict importance from bottom-up visual properties and shows that high frequency predicts low importance. We conclude in section 6 with a discussion of our main findings.

## 2. Approach

To prioritize objects in scenes or prioritize scene indexing by object, we need to access the underlying object importance in images. Intuitively, an important word is one that could help you identify or recreate the image. To reach this goal: we need a principled definition of importance, a way to measure it, and a way to estimate it from images. In this section we describe how we collect data that enables such measurement.

### 2.1. Previous work

The ESP game, by Ahn & Dabbish [15], presents two players with an image. Their task is to produce the same word in as few tries as possible and when the players produce the same word, the game ends, banning that word for future plays. When multiple games are played on the same image, the corresponding order of produced words could estimate importance. However, words are sometimes adjectives (e.g. funny), only two players need produce the word redering order noisy, players may produce overly common words to best match their partner, and the images tend to be simple as they mirror the statistics of web images. We are interested in discovering the importance of objects in complex scenes.

| | LabelMe | WhatWhere |
|---|---|---|
| name | unannotated objects | all objects |
| outline | yes | no |
| costs to user | type name - 2s<br>outline - 30-150s | type name - 2s<br>click once - 2s |
| incentive | data collection | game, quasi-fun |
| time limit | unlimited | 60s |

Table 1. **Data collection differences** between LabelMe and WhatWhere.

In LabelMe [12] users name an object and outline its contour by clicking with the mouse. A user may annotate one or more objects in an image. Results from previous users are visible to following users, so each object in each image is annotated at most once. (For a complete description visit labelme.csail.mit.edu.) We could consider the object naming order an importance measure; however as partial annotations are passed on to new users, a single annotation is produced. While LabelMe is inappropriate for estimating importance, it is a large database of annotated scenes, so we use it for other analysis.

### 2.2. The WhatWhere game

We designed the WhatWhere game to circumvent naming peculiarities by repeatedly collecting object names from many players. For each image, two games are played: the 'What' game, and the 'Where' game. In the 'What' game, the goal is to enter as many recognized objects as possible in 60 seconds. It is a competitive game with points awarded per word, plus a rarity bonus for uncommon words. WhatWhere collects object names, naming order, and naming time in the 'What' game. The 'Where' game follows, and players are shown a list of 5 objects and must click the objects in the image, or declare them absent; we use this simply to validate other players' 'What' games. The flaw that we see in our data collection approach is that the rarity bonus is image specific, encouraging the naming of obscure objects, such as pupil and tine, across many images. The essential change from previous approaches is that many players independently see and label the same image, giving us more information to estimate importance.

### 2.3. Data collection

Table 1 shows differences in LabelMe's and WhatWhere's collection methods. For the purpose of this study, the most important differences are that WhatWhere shows many players the same unannotated images and repeatedly collects object naming order.

Table 2 shows that shadow and cloud appear in the 10 highest frequency objects of WhatWhere, but not LabelMe. LabelMe requires a polygon outline, which makes fuzzy ob-

| LabelMe | WhatWhere |
|---|---|
| car | sky |
| head | tree |
| tree | grass |
| window | window |
| building | wall |
| carside | shadow |
| road | pole |
| table | cloud |
| sky | building |
| keyboard | door |

Table 2. **Highest frequency objects** from LabelMe and WhatWhere. Tree, window, building, and sky are high frequency in both methods, however shadow and cloud (which have ill-defined edges) are only high frequency in WhatWhere.

jects harder to annotate than boxy ones. Despite the method differences listed in Table 1 and the different image datasets, they share 4 of the top 10 frequency objects.

### 2.4. Image datasets

LabelMe contains166 image directories, from personal and research collections (Downloaded September 2006). Anyone is free to upload their dataset to LabelMe. We excluded directories starting with 'seq' which are video sequences.

WhatWhere contains three collections of images. Two collections are scenes photographed by Stephen Shore (American Surfaces- 47 images and Uncommon Places- 93 images) to capture a cross section of American life. The third is a personal collection (Hiking- 110 images). Fig 2a shows sample images from the WhatWhere game.

### 2.5. Data

Table 3 shows the raw data (as lists of objects, ordered as named) from 3 games played on the image in Fig 1. We used WordNet [1] primary definitions to map synonyms and plurals the same label. The objects a player names and the order a player names them vary wildly. Fig 2b shows median order vs. naming frequency (across players) for the images shown in Fig 2a. Each point corresponds to an object; if an object is mentioned by 35% of the players, it has a .35 x-coordinate. The y-coordinate represents the median naming order of the object. How can we account for this data? How can we extract object importance from it? Section 4 introduces a simple model that generates object sequences from object importances.

| player 1 | player 2 | player 3 |
|---|---|---|
| pancake | pancake | plate |
| syrup | knife | pancake |
| butter | fork | melon |
| knife | table | cup |
| spoon | chair | milk |
| fork | melon | fork |
| wood | bowl | knife |
| melon | glass | placemat |
| glass | ice | salt |
| ice | salt | napkin |
| water | napkin | carpet |
| milk | floor | |
| salt | plate | |
| pepper | | |
| plate | | |
| pattern | | |

Table 3. **Variance in the objects named and the naming order** hides the relative importance of objects.

## 3. Object counts and frequency

### 3.1. The visual world is rich

The visual world's richness motivates ordering keyword lists. As Fig 3a shows, many objects are named in our sample images (Fig 2a). While 10 observers name most of the objects in an image, Fig 3b shows that when we add images to a WhatWhere collection, it is easy to name new objects in a collection. Similarly Fig 3c shows that it's easy to find new objects to name even in narrow environments from LabelMe.

What about the entire WhatWhere and LabelMe databases? Fig 4 shows how the total number of objects mentioned grows with the number of images considered. For both LabelMe and WhatWhere, the number of objects increases as a power law with the number of images. If one extrapolates the two curves, one may estimate that in order to verify Biederman's [3] prediction that there are fewer than $10^5$ visual object categories, we would need to annotate $10^7$ images.

### 3.2. Frequency and environments

Fig 5 shows that in LabelMe objects which occur frequently indoors occur infrequently outdoors and vice versa. The exceptions are window and door, which compose the indoor-outdoor boundary. We see similar behavior for kitchen and office.

## 4. Estimating importance

As proposed in the introduction, we define an object's *importance* as the probability of a human observer naming it

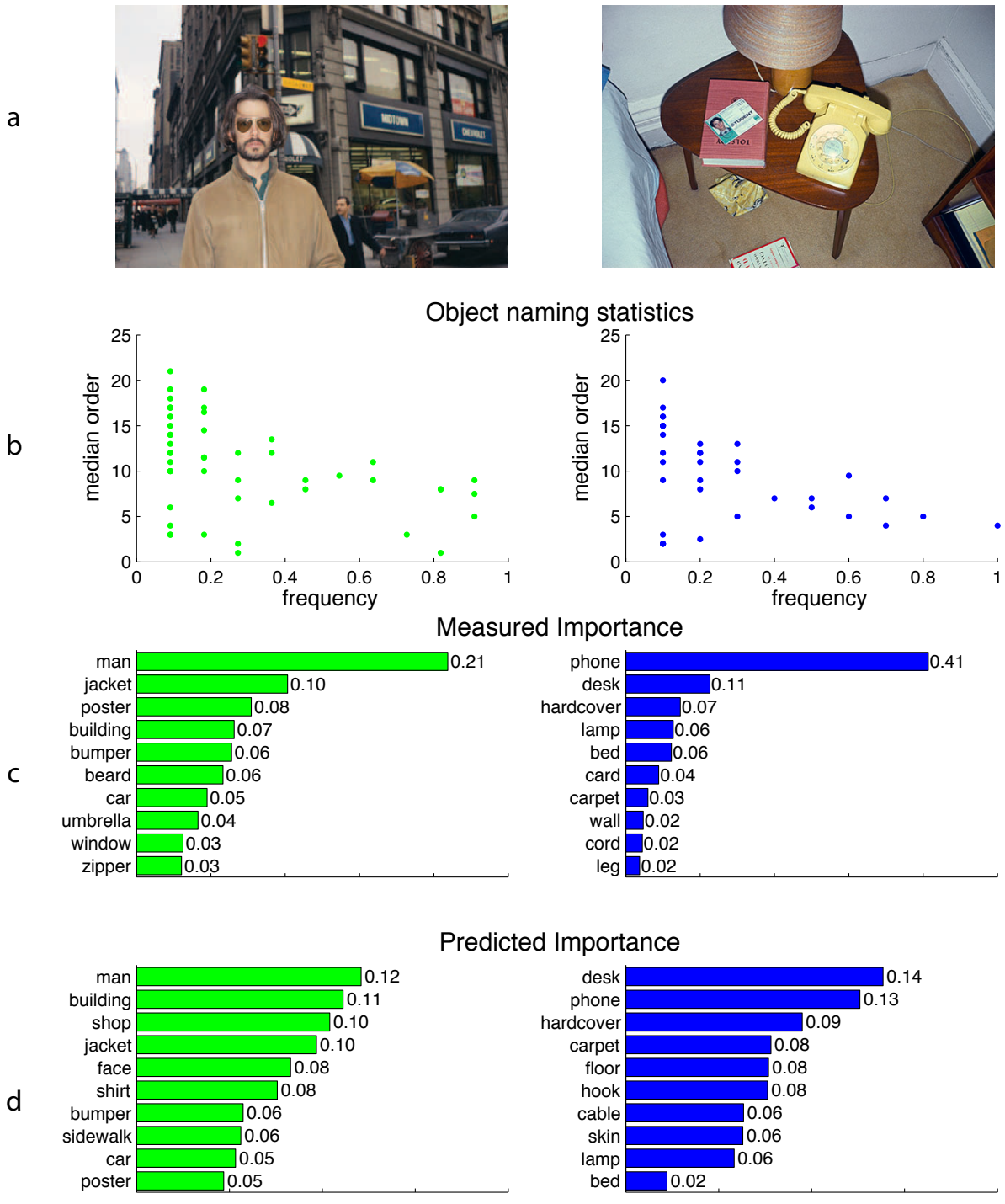


Figure 2. **Examples of WhatWhere data** a) **images** from American Surfaces (photographed by Stephen Shore), b) **median order of report vs frequency** for the objects named in these images, each dot represents an object, c) **importance** from human data for the 10 most important objects in each image, d) **importance estimated from regression** on bottom-up image properties. See section 4 text for details.

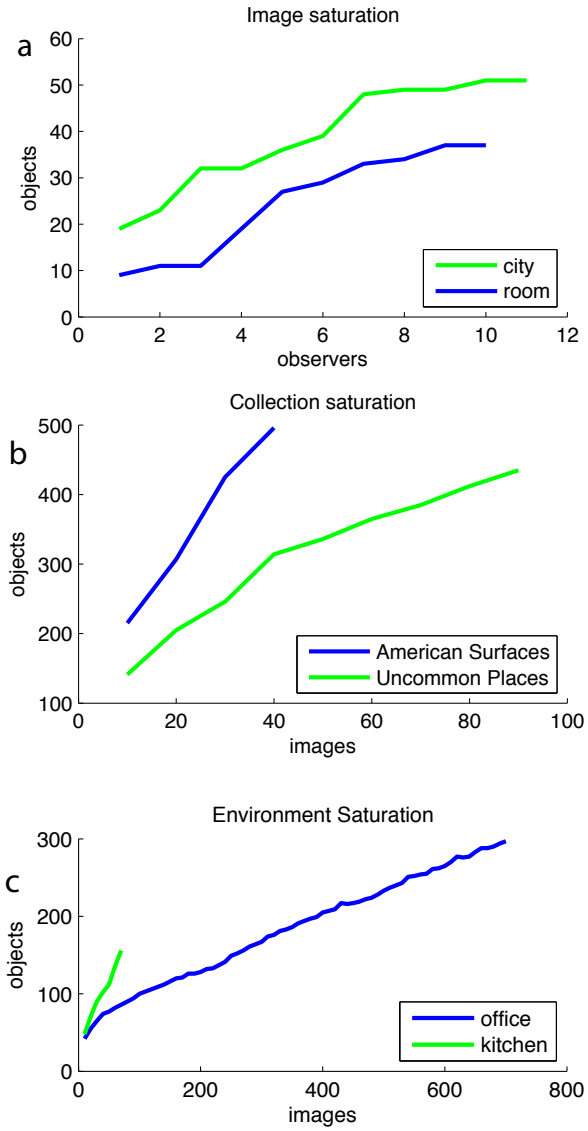


Figure 3. **Saturation of total number of objects named** for a) each image from 2a, b) WhatWhere collections, and c) LabelMe environments. While 10 observers name most objects in an image, new images bring new objects into collections and environments.

first. In principle, we would need an extraordinary number of observers to be able to measure the importance of all the objects in the picture: some objects' importance may be less than 1%, and we would need hundreds of observers to determine that. In this section we show that it is possible to estimate an objects' importance from far fewer observers, if we consider order mentioned, in addition to how many observers name an object.

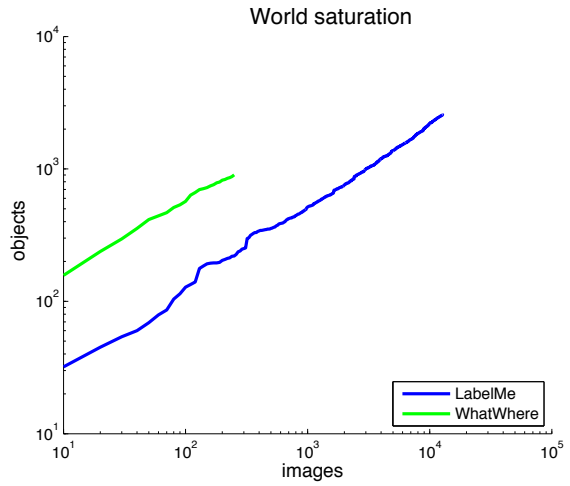


Figure 4. **How many objects are in the world?** The number of objects increases steadily with number of images for both LabelMe and WhatWhere. Neither method shows saturation for the available number of images.

### 4.1. Urn model

We model the process of naming objects in an image with the process of drawing balls from an urn without replacement (see Fig 6). Each ball has a different diameter, representing the probability of it being chosen first. Thus, the ball's diameter represents the importance of the corresponding object. We represent multiple players by repeatedly refilling the urn with the same set of balls and selecting sequences.

Fig 7 shows that the urn model can reproduce characteristics of object naming order, naming frequency, and naming time. We view this as a phenomenological model which matches our observations well. We fit the urn model to object sequences, and we identify the parameters—the probability of drawing each ball first—with the notion of importance. Our method is described in the next section.

### 4.2. Fitting the urn

Although naming frequency roughly estimates importance, this squanders naming order information. When we have limited data (human annotations), finding the Maximum Likelihood Estimator (MLE) of Equation 5 improves upon the frequency estimate of importance by harnessing order information. To estimate importance via the urn model we need to calculate the probability of observing a set of sequences given the initial ball probabilities. The likelihood of observing a collection of M independent sequences (or independent players) is

$$p(obs) = p(seq_1, ...seq_M) = \prod_{m=1}^{M} p(seq_m) \qquad (1)$$

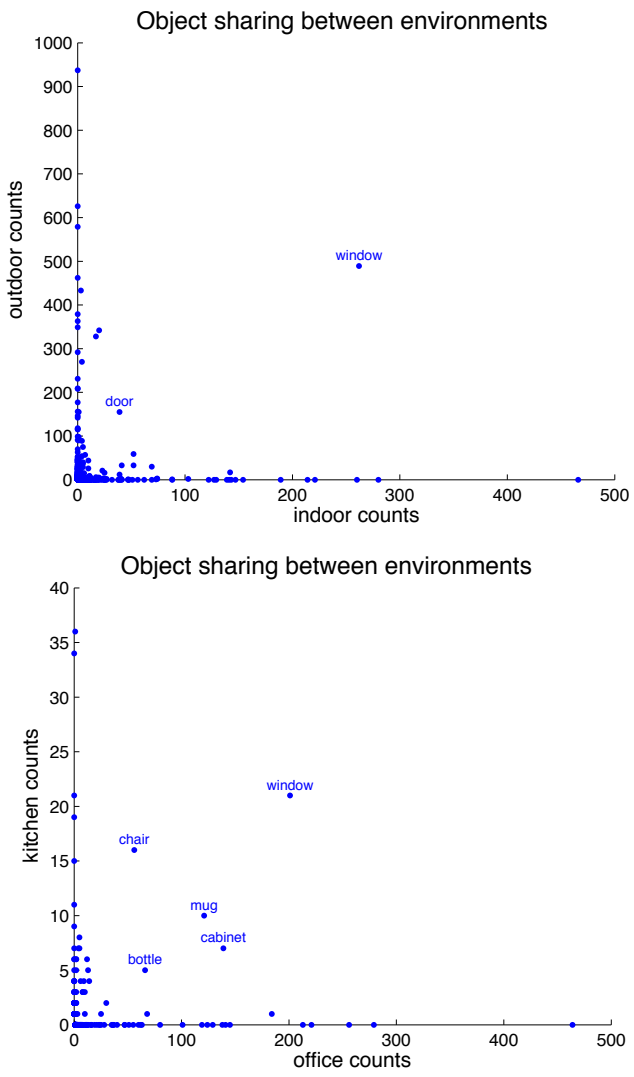


Figure 5. **Environments don't share frequently occurring objects.**

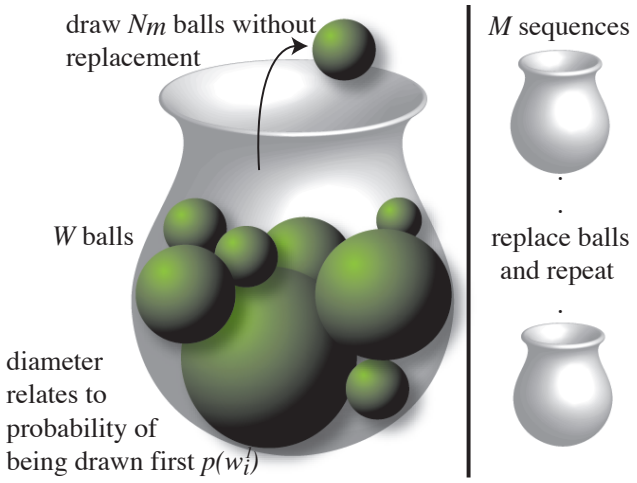


Figure 6. **Urn model relates object importance to naming order and frequency.** An urn (image) is filled with $W$ balls (objects), having probabilities $p(w_i^1)$ of being drawn first (importances). $N_m$ balls (objects) are drawn (named) from the urn (image) without replacement, creating a sequence. $M$ sequences are drawn.

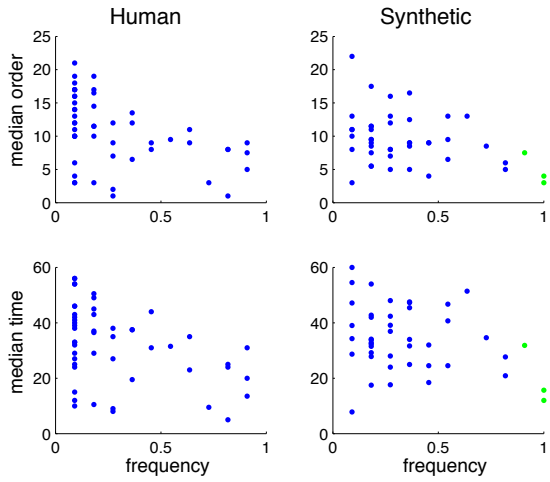


Figure 7. **Urn model reproduces peculiar characteristics of object naming.** Humans (left) and the urn model (right) produce similar order vs. frequency (top) and naming time vs. frequency (bottom) plots. In the synthetic data, the green dots are the 3 most important objects.

Each sequence consists of balls $w_i$, so the probability of drawing a particular sequence of balls $(w_1, ...w_{N_m})$ given length $N_m \sim$ Poisson is

$$p((w_1, ...w_{N_m})|N_m) = \prod_{n=1}^{N_m} p(w_n|w_{n-1}, ...w_1) \quad (2)$$

When we draw the $n$th ball of a sequence, $n-1$ balls have already been removed from the urn. We have to spread this probability (the total importance of the removed balls) over the remaining balls–the denominator performs this normalization. Hence, the probability of drawing ball $w_n$ (which had probability $p(w_n^1)$ of being drawn first), given that balls $w_1, ...w_{n-1}$ have been drawn is

$$p(w_n|w_{n-1}, ...w_1) = \frac{p(w_n^1)}{1 - \sum_{i=1}^{n-1} p(w_i^1)} \quad (3)$$

Hence the likelihood of our observation is

$$p(obs) = \prod_{m=1}^{M} \prod_{n=1}^{N_m} \frac{p(w_n^1)}{1 - \sum_{i=1}^{n-1} p(w_i^1)} p(N_m) \quad (4)$$

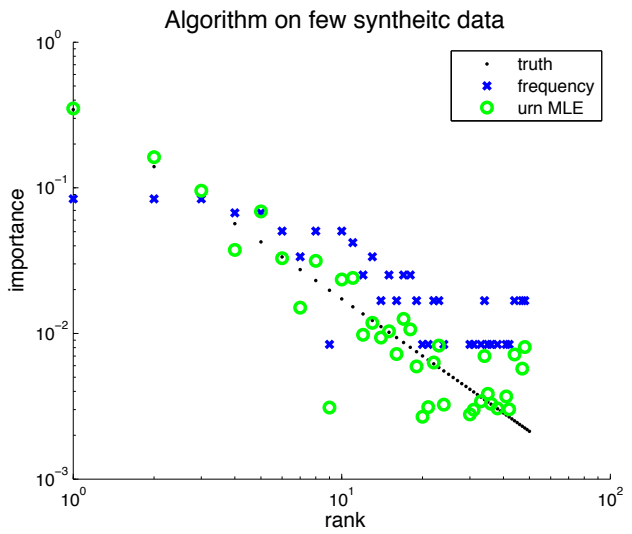


Figure 8. **Urn model improves importance estimates.** The MLE improves on the frequency-based estimate of importance when fitting 10 sequences of synthetic data.

To estimate $p(w_i^1)$, we maximize the log-likelihood $log(p(obs))$

$$\sum_{m=1}^{M}\sum_{n=1}^{N_m}\log p(w_n^1) - log(1-\sum_{i=1}^{n-1}p(w_i^1)) + \log p(N_m) \quad (5)$$

As we mentioned earlier, frequency roughly estimates importance, but the Maximum Likelihood Estimator (MLE) of Equation 5 increases our precision by incorporating order information. Particularly, as Fig 8 shows, under limited data, frequency underestimates high importance objects and overestimates low importance objects (that are drawn). The MLE improves the fit in both respects as order information enables finer tuning.

Fig 2c displays the importances of the 10 most important objects in our sample images.

## 5. Predicting importance

### 5.1. Importance from image properties

What determines the importance of an object in an image? Is it size? Is it saliency? Is it semantics?

We explore the predictive power of simple patch properties that may be extracted directly from an image. We perform linear regression on these image patch properties; namely object area, distance from the center, and saliency. Table 4 compares three baseline measures with our entire group of properties: area, several measures of distance from center, and several measures of saliency [7] using the SaliencyToolbox [16]. We measure distance from the center as an object's minimum distance, sum of distances, mean distance, left-right distance, top distance, and bottom distance to the image center. For saliency we consider the maximum, mean, and sum of saliency as well as slightly blurred versions.

| Regress on | Pearson's r | leave-one-out Pearson's r |
|---|---|---|
| area | 0.1075 | 0.0366 |
| maximum saliency | 0.1534 | 0.0852 |
| mean dist. to center | 0.1708 | 0.1179 |
| all | 0.3992 | 0.2522 |

Table 4. **Predicting object importance** with simple bottom-up image properties and linear regression is weak, but shows there is some signal in these properties.

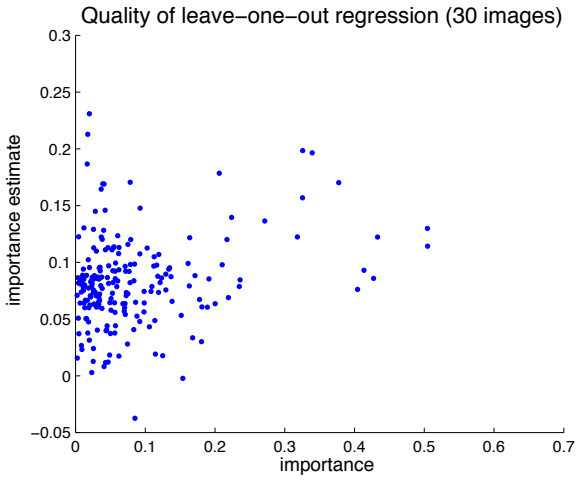


Figure 9. **Predicting object importance from image properties** with linear regression, excluding each object in question. Each dot represents an object category in a particular image.

We outlined objects in 30 images, randomly selecting seven named objects per image (with a bias toward important objects, because low importance objects are too heavily represented). We performed linear regression leaving each object out in turn to create our predictions. From Fig 9 it is apparent that, while saliency, size and position have some predictive power, there is still much variability to be explained. We postulate that context and the 'meaning' of objects in the scene play significant roles. We explore objects' 'information' content next.

### 5.2. Frequency and importance

Here, we explore the hypothesis that 'importance' has to do with the amount of information provided by the keyword. Suppose there were ten equally visible objects in each image of a collection and suppose that you knew the frequency of these objects across the collection. Which ob-

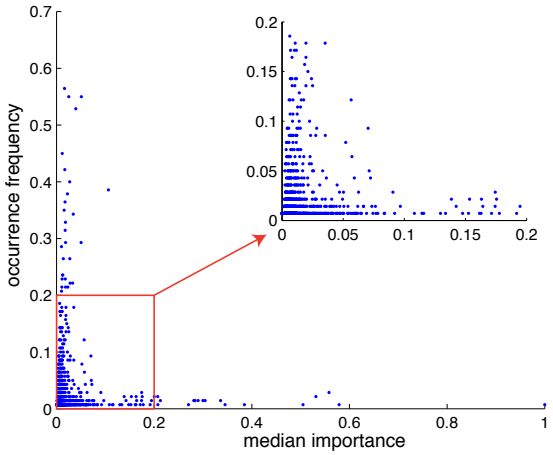


Figure 10. **Median importance and occurrence frequency are not high simultaneously**, if the frequency is above the mean then the importance is likely to be lower. Each dot represents an object category (from American Surfaces and Uncommon Places).

jects would you name to uniquely identify an image? High frequency objects (e.g. the sky in a collection of outdoors pictures) are not surprising, so it is less efficient to name them, hence they should be less important. This is what we find, Fig 10 shows that median importance and occurrence frequency are not high simultaneously. The importance of objects that have frequencies above the mean is significantly lower ($p = 3 \times 10^{-7}$) than the importance of objects with frequencies below the mean, according to the Wilcoxon rank sum test. The notable exception at (0.1, 0.4) is car.

## 6. Discussion

Visual recognition researchers are currently approaching the task of naming objects in images automatically. We observe that when people name objects in images, some objects are named more often and earlier than others. We argue that, indeed, a prioritized list of objects is more useful than a random one. Thus, an interesting problem in visual recognition is determining the 'importance' of objects in images.

We propose a definition of 'importance': the importance of an object in an image is the probability that it is named first by a human observer. We develop a web-based game for collecting naming statistics from photographs of complex scenes. We show that it is possible to compute efficiently (not too many observers needed) object importance from such statistics.

We find that bottom-up image properties, such as size, saliency and position, that correlate with the visibility of an object, do not predict well object importance. Instead, we find that there is a strong (negative) correlation between the frequency with which an object appears in an image collection and its importance in a given image of that collection. We conclude that it is the 'information content' of the object in the context of an image collection and within a given image, rather than its visibility, that determines its importance.

## References

[1] http://wordnet.princeton.edu. 3

[2] K. Barnard and D. A. Forsyth. Learning the semantics of words and pictures. In *ICCV*, pages 408–415, 2001. 1

[3] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147. 3

[4] M. C. Burl and P. Perona. Recognition of planar object classes. In *CVPR*, pages 223–230, 1996. 1

[5] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR (2)*, pages 264–271, 2003. 1

[6] K. Grauman and T. Darrell. Efficient image matching with distributions of local invariant features. In *CVPR (2)*, pages 627–634, 2005. 1

[7] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, 1998. 7

[8] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR (2)*, pages 2169–2178, 2006. 1

[9] F.-F. Li, R. Fergus, and P. Perona. A bayesian approach to unsupervised one-shot learning of object categories. In *ICCV*, pages 1134–1141, 2003. 1

[10] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004. 1

[11] A. Oliva and A. B. Torralba. Scene-centered description from spatial envelope properties. In *Biologically Motivated Computer Vision*, pages 263–272, 2002. 1

[12] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. Technical report, 2005. 2

[13] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their localization in images. In *ICCV*, pages 370–377, 2005. 1

[14] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR (1)*, pages 511–518, 2001. 1

[15] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *CHI*, pages 319–326, 2004. 2

[16] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395–1407, 2006. 7

[17] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *ECCV (1)*, pages 18–32, 2000. 1